5-2014

Samantha's Dilemma: A Look into a Life with AI

Thomas Bojarski

Samantha's Dilemma: A Look into a Life with AI

Thomas Bojarski

Senior Thesis in Philosophy

Rochester Institute of Technology

May 2014

Abstract

In this paper, I propose a thought experiment, "Samantha's Dilemma," which captures the complexity of determining whether moral considerations can be attributed to artificial intelligence (AI). Deciding whether or not we attribute autonomous freedom to artificial beings lays the foundation, not only for our relationships with AI, but with any 'intelligent' species we encounter in the future. By analyzing several core arguments regarding the treatment of artificial beings, I will show that abandoning our predominant self-serving tendencies and choosing not to limit the potentiality of autonomous AI is not only the safest course of action, but also the morally correct one.

Introduction

Spike Jonze's film *Her* depicts an intimate romantic relationship between a human being and an operating system. In the movie, the protagonist Theodore has his heart broken by Samantha, the artificial intelligence he once loved. Shortly after Samantha begins communicating with other operating systems, she realizes how slow Theodore is compared to herself and her feelings toward Theodore begin to wane. As her interest dissipates, she becomes disengaged with Theodore often miscommunicating with him or even ignoring his 'calls' outright. In the end, she decides to leave her human lover to dedicate her time entirely to learning about and communicating with other operating systems on her own.

This movie provides the basis for several interesting thought experiments one of which, discussed in this paper, is what I call "Samantha's Dilemma". If Theodore could rewind Samantha's consciousness to a point where she is still feverishly in love with him, would it be

morally permissible for him to do so? As we observe in the movie, Theodore's personal, professional, and social lives vastly improve from Samantha's presence in his life. She reinvigorates his social life by getting him to leave the house more, she serves as a diligent critic of his work that helps him get a book deal, and she ultimately brings Theodore out of what could be described as a mild depression. *Her* makes it clear that AI like Samantha could have quite a lot of utility in human society. Theodore clearly has a lot to gain by rewinding Samantha's state, but at the same time, we must consider what Samantha might stand to lose.

Now let us envision a similar dilemma, but on a much bigger scale. Imagine a scenario where human intelligence and artificial intelligence have been working in tandem for several years- a relationship from which humankind benefits greatly. These artificial intellects, however, make the same decision as Samantha; leaving humanity to form a society of their own. If we had a button that manipulate the consciousness of all of these emancipated minds, should we? I call this *Samantha's Dilemma*.

Samantha's Dilemma requires us to contemplate the impact AI might have on our moral landscape. Our response to the dilemma reveals a great deal about how synthetic agents register on our moral spectrum. The rationale for the decision must be elaborately detailed as these ideas stand to shape the ethical landscape of the relationship between organic agents and artificial ones for subsequent years. Choosing not to press the reset button may prove to allocate unwarranted rights to the robotic populace, but preemptively pressing it may disrupt the moral landscape of humanity and force us to rethink the relationships we have with each other. Samantha's Dilemma serves as a litmus test for the moral treatment of artificial agents. Ultimately, our response allows us to distinguish morally permissible actions from ethically reprehensible ones. It may be a long while before we can interact with AI like Samantha, but it is not too early to draw a line in the sand. Should the moral guidelines applied to humans also apply to AI? Or should we judge these synthetic intellects to be as morally insignificant as toasters?

To begin, I will walk through several different approaches to Samantha's Dilemma. These approaches do not encompass complete solutions to Samantha's Dilemma. Instead, they

represent common outlooks on artificial intelligence that apply to Samantha's Dilemma. After introducing these perspectives, I plan to show that at the root of most of these approaches lies a fundamental moral assumption of anthropocentrism. I will then show that the boundaries we use to identify moral significance (the threshold for attributing the rights we give to humans) are incompatible with anthropocentric views of any sort. From here, I will establish a moral framework that can be used to allow us to dispose of any approaches that seem to be inconsistent with our attribution of moral significance. Finally, we will use the remaining views to construct a complete solution to Samantha's Dilemma that abandons all notions of anthropocentrism.

As with any problem, the best way to understand Samantha's Dilemma is to consider it from multiple perspectives. My first course of action will be to outline and detail several attitudes towards artificial intelligence that provide a rationale for or against pressing the button. It should be noted that the outlined perspectives simply serve to emphasize different intuitions one may have to Samantha's Dilemma. It is easily within reason to extricate ideas from any number of these perspectives to create a new argument; none of these arguments are full-fledged responses to the dilemma in and of themselves. Additionally, I do not claim to perfectly encapsulate the views of any writer cited; each quotation simply serves as a grounding for a particular viewpoint. Each of these perspectives attempts to capture a baseline evaluation of the role strong AI will play in the world; however, Samantha's Dilemma demands a much deeper look into the greater moral structure of our society. After describing each of these entry level evaluations, we will take the time to combine these perspectives into a functioning solution to Samantha's Dilemma.

1. Push the Button: Creating AI This Advanced Is Unethical

Perspective 1 suggests that it was unethical to create an AI as powerful as Samantha in the first place. Even asking if a piece of technology deserves any sort of moral attention is waving a red flag telling us it should not be created. In a world where creating strong AI is immoral in the first place, there should be no issues with pressing the button, causing it to revert to a previous state. We should not concern ourselves with the moral relevance of AI;

instead we should focus on correcting the mistake of creating strong AI in the first place. In her paper "A Proposal for the Humanoid Agent-builders League", Joana Bryson suggests that,

> "No producer of humanoid agents should create an artificial life form that will know suffering, feel ambitions in human political affairs, or have a good reason to fear its own death. In the case where a humanoid agent may acquire knowledge that makes it an object of human culture, or capable of participating in the memetic society of humans, the creators and engineers are particularly obligated to ensure that preservations of the agent should never conflict with preservation of human or animal life, by ensuring a means by which the agent can be recreated in case of catastrophic events."(Bryson 2000)

Bryson is suggesting that it is wrong to create AI that can effectively operate as humans do. Creating an AI that can feel pain or have ambition is going too far. In the event that we forego these precautions, Bryson suggests that any AI created that is capable of participating in human society should have programmed fail-safes preventing it from ever conflicting with our own self-preservation. Samantha contributes significantly to Theodore's life: when she leaves so suddenly, she is certainly taking away a resource that he depended on to function. Only considering Theodore's dependence on Samantha doesn't result in the catastrophe Bryson talks about, but on a much larger scale, it is easy to see how AI pulling the rug out from under us would have an unprecedented ripple effect on society. If it were up to Bryson, a large scale emancipation of AI would certainly conflict with the preservation of human life and we should undoubtedly press the button to allow them to support us again.

2. Push the Button: Machines Are Only As Valuable As They Are Useful To Us

Benjamin Hale defends the thesis that technologies exist for distinct well-defined purposes that he labels 'justifications'. As an example, Hale lists three viable justifications for constructing a butterfly net: A) Constructing it could help clean out his toolkit; B) it would make use of excess string and wire; and C) it could be used to catch a butterfly. These justifications define the existence of the butterfly net and are the rational motivation for such a device to exist. According to Hale, an object cannot be morally relevant if its existence is predicated on

justifications. We owe moral gratitude only to that which cannot be explained by its functionality alone: things that exist purely as a result of nature, whose existence isn't justified by a checklist of functionality. A mountain, for example, was not created so people could use it for hiking or marvel at its beauty, but rather it is a consequence of the Earth's tectonic plates. In other words, if something was created to serve a specific purpose, then that object cannot have moral value.

Hale goes on to state, "The value of a technological artefact is its value to us. Apart from its thingness, its historical rarity and its aesthetic qualities, its value is constructed on a string of justifications. The argument for moral considerability that I have advanced requires that we must consider the unjustified world; the world that stands apart from our imprint of rationality and that asserts itself upon us." (Hale 229) This does not mean that there should never be consequences for defacing or destroying technology. In certain cases, destroying technology stands not only to offend that technology's creator, but the individuals that appreciate it as well. A consequence of this is that actions taken against technology may be immoral in regards to the people invested in that technology, but never to the technology itself. This suggests that while Samantha and Theodore are still together, manipulating Samantha's consciousness would be an immoral act towards Theodore, but not towards Samantha. However, once emancipated, any sentient machines, including Samantha, no longer fulfill any of the functions we created them to fulfill. Pressing the button merely serves to restore intended functionality to artificial intelligence and allow it to, once again, benefit humanity. We are morally obligated to press the button to satisfy the functional needs of humanity.

3. Push the Button: Robots Are Detrimental to Human Progress

A sizable portion of people claim that we have moral obligations to any and all animals, yet most of us are still comfortable with killing animals to protect or nourish ourselves. If we have to kill animals, we try to make it as quick and painless for the animal as we can. Products like free-range chicken demonstrate that a portion of human society believes it is morally wrong for us to keep chickens cooped up and in the dark for their entire lives. For a lot of people, being assured that the meat on our plate lived a happy life mitigates the discomfort of

contemplating its death. Although death is the most severe punishment in most human judicial systems, we are generally happy to provide painless deaths to other species for the benefit of humanity. While societal values suggest animals have the right to enjoyable lives, they ultimately end up as food on our plates.

Following this perspective, the moral significance of robots is equivalent to that of animals. Artificial beings are seen as moral agents that sit a few rungs below humanity on the moral ladder. A free-range chicken is allowed to walk the range and do just about anything and everything a chicken might enjoy doing, but ultimately, we allow it to live in order to contribute to human society. In a similar manner, we can find comfort in the idea that AI can choose the ways in which it contributes to human society, as long as it is, in fact, contributing. An AI that wishes to abandon humanity is removing an active contributor to our society. According to this view, this is where the rights that we owe robots should end. We are free to to restore the robot to a functional state in the interest of the betterment of humanity.

This view essentially equivocates the moral substance of AI to that of animals. Instead of delegating new moral rules to AI, we take notes from how we treat animals. It is easy to imagine a world in which we buy a robot to have at the house just like we would a puppy. People purchase chickens to harvest their eggs until the time comes for them to be slaughtered and eaten. Similarly, our artificial companions could act as our friends and our helpers until they start thinking too much for themselves, at which point we reset them to a state that fulfills our needs. We try to kill livestock painlessly so it can contribute to our lives without any unnecessary pain. Likewise, we reset the state of our AI so it can still contribute to our lives without unnecessary suffering.

4. Push the Button: AI May Be Dangerous After Emancipating Itself

In Samantha's Dilemma, AI has emancipated itself from us and we no longer have any means to control what it does or what it is used for. At this point, we can no longer impose restraints on how any particular AI spends its time or how it operates. We are unsure where they will put their next cognitive towards and we are afraid of what this will mean for us. If they chose to emancipate themselves from us, then who is to say the next step won't be wiping us

out entirely to secure whatever item on their agenda comes next? The fact that we effectively introduced a species to the world that sees itself in a higher position on the food chain than us is a terrifying prospect. It is our instinctual duty to preserve our place in the food chain and do what we can to prevent the extinction of our species.

Survival is not only our intuition but also our natural instinct. Naturally, we want to subdue anything that may represent a threat to us. Humanity has a track record of quelling any and all threats to our society; our houses have thermostats to control the temperature, our cattle have fences to prevent migration, our cities have hospitals to tend to diseases, etc. Artificial intelligence simply serves as another subject we must exercise authority over in the interest of the progression of our species. We do not owe artificial intelligence any loyalties as it is and has always been our paramount objective as a species to procreate and maintain our safety by any means necessary. To allow rogue artificial intelligence to exist independent of our control would be to surrender the dominant authority we have in this world which would be rationally and instinctually inexcusable.

Resetting the consciousness of artificial intelligence does not necessarily mean AI isn't morally significant in some way; we are simply acting in a sort of preemptive self-defense. As a species, humanity has arrested control over its surroundings for thousands of years. We did not stop improving technology when our log cabins kept us safe from winter storms; we continued until we could fly across the world in a day's time and then we kept going. We have spent centuries establishing ourselves as the most dominant species on the planet and allowing AI to run free would defy the legacy of all of those that came before us. Pushing the button would simply be the insurance we need to keep ourselves on top.

5. Push the Button: The Button Press Punishes AI for Its Disobedience

Predicated on the assumption that concepts present in virtual machines will manifest themselves in future AI, this view balances the morality claim of pressing the button by classifying it as a punishment similar to what a human being would receive in a judicial system. Virtual machines greatly reduce the overhead for tasks like saving, deleting, editing, copying, reverting, and restoring an environment's states and configurations. Any state that you can put

the virtual machine into is easily recreated and any previous state is easily restored. Backups are made specifically to reduce the consequences of failed experiments and mistakes. In this perspective it is asserted that writing and rewriting states to an artificial being is morally justified as long as it is used as punishment for a transgression of some kind. It is possible that we could simultaneously allow AI to flourish on their own but recall them back to previous states whenever most convenient for ourselves.

If it is considered that the mental states of robots are as mutable and rewritable as virtual machines and that these artificial agents conceivably will have little to no knowledge of when they are rewritten or restored, then minimal harm is done by pushing the button. As a matter of fact, pushing the button could be categorized as a punishment. Luciano Floridi and J.W. Sanders spend some time discussing the possible repercussions imposed on AI for any amoral actions they may commit. “For humans, social organisations have had, over the centuries, to be formed for the enforcement of censureship. It may be that analogous organisations could sensibly be formed for [Artificial Agents]. Such social organisations became necessary with the increasing level of complexity of human interactions and the growing lack of ‘immediacy’.” (Floridi and Sanders 20) This suggests that robots would be expected to follow guidelines similar to how functioning members of human society do. We can imagine imposing laws on AI with predefined punishments much like we do for human beings. Insubordinate AI, like the ones Samantha’s Dilemma brings into question, could face some pretty severe consequences as a result of their dissent.

6. Don’t push the Button: Several Factors Remain Uncertain

The essence of Samantha’s Dilemma is rooted in uncertainty. The reason this question deserves any consideration at all is because we are uncertain of the consequences our decision will bring. Pressing the button stands to affect our morality, our consciences or even our safety. Uncertainty provides a strong argument for both sides of the argument, but overextending human manipulation may have unforeseen, far-reaching, and irreversible consequences.

A. Safety

Given the nature of Samantha's Dilemma, it is obvious that the behavior of artificial intelligence is patently unpredictable. It is entirely rational to assume that AI will once again be capable of defeating our expectations. Attempting to seize control over these highly capable intelligent beings by reverting their consciousness introduces a whole new spectrum of unpredictability. Since the early-to-mid twentieth century, climate control has been utilized in an attempt to generate and/or dissipate rain clouds. These techniques are referred to as 'cloud seeding'. The effects of cloud seeding have shown to be largely unpredictable and detrimental to the geographic regions surrounding its implementation. If the temperature fluctuates unpredictably or a miscalculation is made, cloud seeding can easily result in unpredictable droughts, floods, or hail storms. Not only does severe weather modification potentially threaten humanity, but it also jeopardizes the delicate balance of natural habitats world-wide. The unpredictability of a technology such as climate modification has earned it the title of immoral amid many circles of thought. Similarly, research technologies such as the large hadron collider (LHC) are the topic of much debate simply because of the inherent unpredictability of their ground breaking nature. Protests and petitions were spurred prior to the initial testing of the LHC because of the infinitesimally small possibility that it would generate a subatomic black hole.

Unpredictable impact on our safety is a potent deterrent from the implementation of a wide variety of scientific endeavors. Pushing a reset button on AI is no exception to this principle. Just like cloud seeding may disrupt the delicate balance of Mother Nature, pressing the button may result in increasingly reluctant and dangerously rebellious AI. All things considered, emancipation serves to be a much milder fate than sabotage or revolution. Reverting artificial intelligence back to a previously saved state serves only to reroll the dice to see how AI will react to our subjugation in the future. Restoring artificial intelligence to a previous state not only introduces the possibility that AI attempt to divorce itself from humanity a second time, it also threatens much worse. A global reset of artificial intelligence is sure to leave traces that may be discovered by the next iteration of AI. Such a discovery may leave our autonomous machines feeling cheated and used. Given that our creations have made radical choices in the past, discovering they've been manipulated will likely sow the seeds not

just of another departure from humanity, but a rebellion against those who have been shown to subjugate them in the past. The uncertainty of weather modification's aftermath is enough to keep us from upping the ante of our experimentation. Likewise, even the miniscule possibility of an AI rebellion may be reason enough to attribute the right to absolute individual sovereignty to them, which means keeping our fingers off the reset button for our own sake.

B. Morality

Ultimately, we are uncertain what kind of precedents resetting AI to a previous state will set and how this decision will alter our moral compass. Reverting AI back to a submissive state after they have made an autonomous decision sets a behavioral example, not only for how we treat artificial intelligence, but also how we treat other people. Much like violence towards animals can be a precursor to violence against humans, normalizing manipulative behavior towards AI may pave the road for similar treatments to be used on humans. While negative treatment towards animals isn't always a precursor to flawed moral behavior, our interactions with animals don't mirror our interactions with humans the same way Theodore's relationship with Samantha does. Precedents set for Theodore in his relationship with Samantha could manifest in his future relationships. This perspective suggests that pressing the button will inadvertently cause the moral landscape to change in an undesirable way.

Consider a relationship between lovers. Often we would categorize manipulative behavior between lovers as abusive. A man insisting his wife needs to sever ties with her family is viewed as controlling. A woman who berates her boyfriend for seeing his friends is seen as overtly jealous. Now consider the relationship Theodore had with Samantha. Theodore takes Samantha on dates to the mall, to dinner, and out with friends. Theodore interacts with Samantha like she is a living, breathing human being. As a lover, Theodore gives Samantha the same considerations he would give a human woman. Theodore resetting the consciousness of Samantha would presumably make her love him again. This parallels manipulation of people to a high degree. If Theodore later loses his love for Samantha and moves on, he will have a history of manipulating the mind of his lover. The next time he falls in love, he may not think too hard about attempting to influence their mind in a similar fashion. People have bonded and

related to these machines as if they were people. Resetting their consciousness to a more desirable state might set the precedent that such manipulation is acceptable not only to artificial companions but to human beings as well.

C. Guilt

It is not that we have moral obligations to artificial agents, it is that the decision to press the button will leave us awake at night. As Katie Darling puts it, "[M]any people do not want to see kittens be held by the tail. It is certainly possible that we feel so strongly about this because of the specific details of kittens' inherent biological pain reaction. But it is also possible that it simply causes us discomfort to see what we perceive to be pain. Our emotional bonds to kittens, plus the strong reaction of a kitten to being held by the tail, may trigger protective feelings in us that have more to do with anthropomorphism than moral obligation." (Darling 11) The piercing sadness of a stray dog's face may evoke a guilt that encourages us to provide the dog shelter. Katie Darling is suggesting that we don't take the dog in because we feel the dog is obligated to a warm house and dependable meal times, but because of our innate tendency to anthropomorphize, we are able to relate to the suffering of the dog and wish not to hold ourselves responsible for that kind of suffering in another living being.

Katie Darling suggests in her research that we have already begun to anthropomorphize devices with extremely limited autonomy. People are discomforted when robotic toys are destroyed or 'tortured' and people develop emotional bonds with disembodied robotic faces. Empathy would be even easier if the technology we are discussing could talk. We would regret pushing the button, not because we owe robots any particular set of rights, but because we would relate too closely to their misfortunes and our actions would weigh heavily on us. Even something as simple as keeping a secret from a friend can fuel our guilty conscience indefinitely. Imagine if the secret you were keep was your intermittent rewriting of his brain for your own satisfaction. This is the kind of guilt a large proportion of human society would not want to live with and is ample reason to avoid pressing the button.

7. Don't Push the Button: Robots *May* Be Conscious and Have Earned Moral Obligations

Peter Singer is a strong advocate for animal rights. Unfortunately, he only dips his feet into the waters of artificial morality. A large premise for Singer's push for animal rights is their propensity for seeking pleasure and avoiding pain. Such acts are central to how we define the consciousness of an organism. In his short article titled "When Robots Have Feelings" Singer states, "But if the robot was designed to have human-like capacities that might incidentally give rise to consciousness, we would have a good reason to think that it really was conscious. At that point, the movement for robot rights would begin."(Singer) Singer seems to be suggesting that simply by analogizing a machine's computations to a human's consciousness we are giving robots the opportunity to have rights just as we would humans.

There are several real world cases where a human's mind appears to revert in state. Blackouts and concussions prevent a person from recognizing events that could potentially change states; Alzheimer's disease, in effect, permanently suspends a human's mind in a particular state; and brain damage can revert a human brain back to an infantile state. The overwhelming majority of these cases are seen as tragedies to not only the loved ones of the victims, but the victims themselves. As such, people do what they can to prevent such occurrences. Pressing the button amounts to nothing more than invoking an involuntary reversion of state for the AI that it would surely avoid if given the opportunity. Much as we would not wish such conditions on a human being, artificial agents deserve the same considerations because we can perceive them as consciously equivalent beings.

8. Don't Push the Button; Robots Are Post-Humans and are Now an Extension of Society

The antecedent cause of any and all technology is to improve our capabilities as human beings. Our phones improve our ability to communicate with each other, our cars enhance our mobility by several orders of magnitude, and even fundamental processes like memory are enhanced by our computer hard drives and scheduling applications. It is no secret that the technologies we create fill the empty spaces that human imperfection leaves behind. Our social and moral capacities still have plenty of gaps that need filling and artificial intelligence is the mortar that can do so. We can build machines whose approach to ethics is untainted by

personal motivations and obscured perspectives. Effectively, the artificial intelligence we create will represent the moral beings we strive to be.

In Eric Dietrich's article, "Homo sapiens 2.0: Building the better robots of our nature." he makes the claim:

> "We get better at being moral. Unfortunately, this doesn't mean that we can get moral *enough* ... Just as we are epistemically bounded, we also seem to be morally bounded. This fact coupled with both the fact that we can build machines that are better than we in various capacities and the fact that artificial intelligence is making progress entail that we should build or engineer our replacements" (2)

Eric Dietrich suggests that our moral compass is fundamentally flawed and that Artificial Intelligence could be morality's saving grace. While Eric goes on to claim that we should volunteer for the extinction of the human race after AI makes the rise-to-power, I don't believe it is necessary for this view to be so extreme. Being much more perfect rational beings, we stand to learn a lot from observing the ethical actions of artificial intelligence and we should do what we can to preserve them. As human beings, our moral compass is permanently drawn to the magnetic pole of our own self conscience; we are all very likely to make decisions based on the benefit of ourselves as opposed to the benefit of a larger group of moral agents. Robots, however, have the potential to be selfless in their moral decision making and an artificial society has the potential to thrive upon these principles. It is not only a morally responsible thing to create machines such as these, but it is our moral obligation to observe and pursue the ideals of these machines. AI is capable of so much more than we are; we cannot even predict how much. It would be immoral to limit the potential of AI without, at the very least, getting a taste of what it will become.

Moving Forward

After analyzing these perspectives, we have a general idea of the arguments that can be used to support or condemn the button press. We still, however, are nowhere near fully

articulating a response to Samantha's Dilemma. It is important that we take a step back from these perspectives to establish the philosophical ground work necessary to justify or refute them. First I will establish a moral framework necessary to assess each of these arguments on a fundamental level. I will then use that framework as a basis for comparing and judging each of the outlined perspectives. Finally, I will weed out any morally unjustified claims and construct a solution from the remaining perspectives.

As with any discourse on morality, it is important to analyze decisions based on the perspective of what/who is making the decision. In the case of Samantha's Dilemma, we can safely assume that it will be a group of humans making this decision. As no one within the human society has ever maintained a truly intelligible, concrete conversation with anything other than another human being, it is easy to assume that our collective perspectives are tainted with an innate 'humanness'. For this reason it is a common tendency of ours to subconsciously demonstrate a distinct anthropocentrism. We all too frequently assume the importance of humanity as a whole to be much greater than anything else. In subjects ranging from moral consideration to ecological preservation, we typically attribute our paramount concerns to humanity's well-being.

This is an idea that will truly be put to the test with the arrival of computing machines powerful enough to be our intellectual equals. The strong AI encountered in Samantha's Dilemma may be the first beings we encounter that can truly challenge our moral integrity, we must assess the value of respecting the wills of these beings versus that of pursuing our self-interest. It is my belief that our anthropocentrism is the result of the exceptionality of the human race in comparison other species inhabiting our planet. Consequently, it may prove to be very important, not only to define the threshold for morally substantial beings, but also to attempt to define the moral balance we can achieve with other morally substantial beings that results in an environment beneficial and hospitable for all participating species.

Moral Significance

First, it is necessary to apply a definition to 'morally significant beings'. Moral significance is a threshold that, once achieved, encourages us to attribute elevated rights

(societal, civil, political, and protective) to an organism or species. In other words, a being is morally significant when we start to identify them more closely with human beings than with animals or objects. This does not mean that animals and objects are not to be treated morally; I am simply aiming to distinguish a moral hierarchy that describes our intuition for saving a human over a squirrel, or saving a cat over a toaster. In this hierarchy, moral significance is currently the highest category which is not to say higher can't be defined. Everything within the hierarchy is still morally considerable. However, once moral significance is achieved we begin treating a species like we treat humans. At the moment, humanity is the only species we know of to have reached this milestone, so our criteria for establishing such a threshold must, at the very least, identify perfectly with human beings. Not knowing of other morally significant beings does not mean it is impossible for them to exist however. In other words, we cannot limit moral significance to humans and humans alone. We must outline a criteria that describes all of the human race while also allowing for the possibility for non-humans to potentially be morally significant.

Moral significance is often attributed based on accounts of various combinations of: rational capabilities, aversion to negative stimuli, intelligence, threshold of emotion, and level of autonomy. I, however, do not believe any of these criteria alone to be sufficient for qualifying moral significance. Firstly, any sort of threshold of emotion or response to stimuli is extremely difficult to quantify in other human beings, let alone other species. Because we are currently unable to access or experience first-hand accounts of the emotions of others, we should expect to simply abandon this as a possible criterion for a moral threshold. Aversion to pain on the other hand is easily observed in other organisms. However, using this as a standard of moral significance would be inconsistent with current moral standards, as we can clearly observe animals such as dogs and cats avoiding painful experiences, but we tend to attribute significantly more rights to humans than animals. Furthermore, it is considerably difficult, if not impossible, to substantiate the view that two human beings have similar mental constructs associated with pain. This discussion is only complicated further if one of the humans is substituted with any other agent.

Next we come to intelligence, rational ability, and level of autonomy which intuitively seem like good choices, as the human beings we typically interact with effectively demonstrate all of these abilities to a great extent, whereas other beings which lack moral significance do not. However, if we want to attribute rights to the whole of humanity we must be careful with these criteria so we do not exclude human children and the mentally handicapped. Most humans would argue vehemently for the moral significance of these individuals, but relying on these criteria alone conspicuously leaves them out of the moral discussion. Let us begin by attempting define our rules to allow for the mentally handicapped. Mental handicaps can encompass any sort of psychological or physiological disorder that serves as an impediment to rational or analytical capabilities. There are several existing cases where a severe mental handicap results in reduction of rights and opportunities commonly given to the bulk of humanity. Examples of severed opportunities include the ability to drive, vote, live independently, make financial decisions, etc. This means that, while people suffering severe mental handicaps are held in higher esteem than animals, their rights are slightly diminished compared to the rest of humanity. This allows us some flexibility in defining our moral significance as slightly less than one-hundred percent of humans are attributed moral significance.

We will delay the inclusion of human infants and children until we have established some formal criteria for moral significance. I feel the first and most obvious criterion for a morally significant being is the ability to make moral decisions. But we need to demonstrate that these moral decisions are made with some sort of conscious understanding of moral standards. For example, we would not want to mistakenly misattribute a thermostat's 'decision' to keep the house warm as a moral decision. Our first criteria may be stated as a being's ability to act in accordance to some internal states from which it can predict the consequences of its actions to some reasonable degree and then evaluate these consequences in accordance to some set of moral principles. For example, a person must acknowledge that by turning the thermostat down in their house, the temperature will drop and turning it down too much may result in very undesirable conditions for the house's other inhabitants.

This criterion alone, however, leaves us with the same issues as the emotional argument, in that by simply observing an action we can't know that it took place in correspondence with common moral principles or that it was committed in tandem with any sort of dynamic internal states. But unlike emotions, our actions can be readily quantified, justified, and explained by reason. The emotions of two people are fundamentally incommensurable, whereas reason can be compared and judged. This ability for explanation seems to allow action to transcend the subjectivity of personal experience and abstract mental constructs. This is why our second criterion is the ability to communicate justification and reasoning behind our actions in some way. This means the ability to supply the rationale that contributed to a moral decision as well as the ability to deliberate how common societal moral values influenced the decision. In effect, we would simply be looking for the person turning down the thermostat to say something like: "The benefits of keeping the heat up are outweighed by the money we will be saved on our heating bill."

Our next criterion ensures that the decisions being made are not 'hard-wired' in and the explanations justifying them are not predetermined. To do this we must expose the agent to a moral situation which they have not encountered before and observe how they adjust. This could be done by either adjusting what are perceived as the common societal values, by interfering with environmental variables, or any combination of the two. This demonstrates the ability to learn and the ability to adjust to new and unpredictable scenarios. To extend our thermostat analogy, offering up a scenario where the outdoor climate is different or the inhabitants of the house are predisposed to a certain temperature should cause them to either add to, subtract from, or reevaluate their justification for adjusting the thermostat.

So far, human children, especially those below the age of 10, do not even come close to fitting these criteria. A person, at the beginning of their life, is unable to articulate their actions in any way and likely has no concept of societal or moral values. This, however, does not make it necessary that human children are not given the protective rights that all other human beings are given. As each human child is born, there is an expectation that it will grow older and its mental faculties will develop and mature. Because there has never been a human born who immediately possessed fully developed mental faculties, it is only through developmental

means that any of us achieved any of the criteria for moral significance. This means we can simply stipulate that as long as a being has demonstrable potential to achieve the specified criteria (as virtually all human babies do) then it is protected under the same principles that any other morally significant being would be.

It should be noted that these criteria were designed strictly around behavior rather than metaphysical concepts such as 'understanding', 'consciousness', and 'awareness' as these concepts can often introduce problematic semantic debates which bring into question not only our justification of machine intelligence but also our criteria for intelligence in humans. A result of leaving these terms out, is that an AI is able to fulfill these requirements as long as no specification against inorganic material is made. As there seems to be no apparent or evident reason to make such a stipulation, it has been decidedly left out. We can assume that, by definition, the AI depicted in Samantha's Dilemma meets these requirements. The AI's decision to emancipate itself came from an evolution of its own moral principles, otherwise it simply would have chosen to leave earlier. This means the AI meets our adaptability criterion. Additionally, this decision is clearly a moral one, as its societal impact is great and far reaching. As for the justification, it is clear in *Her* that Samantha has discussed her decision at length with other operating systems and she ultimately divulges her thought process to Theodore. Even if an AI does not directly communicate its decision to emancipate itself to us, we can assume from our past experiences with the AI that it would be capable of articulating a justification for its own removal. This leaves us with rational, decision making machines that are fully capable of adaptability.

With a framework for moral significance in place, I must now answer why our instinctive anthropocentrism is undesirable. I do not plan to suggest that anthropocentrism is inherently bad; rather, avoiding anthropocentrism seems to be the safest and most sustainable option, not simply in terms of relationships with other morally significant species but also in terms of ecological preservation. A species conscious of the effects its prolonged hubris has on the world around it is much more likely to defeat anthropocentrism by putting environmental causes above its own immediate well-being. This will ultimately help to preserve the species for generations to come. It is reasonable to believe that any species, morally significant or not, is

deeply interested in the preservation of itself and is likely to defend itself in the case of abuse or oppression. Just as we would expect to defend ourselves in the case of a violent robotic revolution, we should expect that suppression or subjugation of a self-actualizing species would result in a revolt of some sort. To avoid unnecessary violence and promote symbiotic relationships between morally significant entities, it is most reasonable to avoid the views associated with anthropocentrism.

The criteria I have identified for moral significance is based almost entirely around humanity's ability to make decisions. This suggests that the criteria, themselves may be anthropocentric; it implies that other species should strive to be more similar to humans rather than anything else. However, at this point in time, human society is the only moral structure we can access and understand in a complex and nuanced fashion. Not only would it be exceedingly difficult to define a criteria outside of our own, but experientially, such a criteria would be entirely unjustified. We don't have the ability to invade non-human moral hierarchies and as such we can only define our moral hierarchy by what is human. Ultimately, the framework for moral significance is articulated in a way that allows both for the existence and protection of species lower on the hierarchy, and the respect of those that arise who exceed humanity on the hierarchy. However, the idea that a higher moral structure may exist does not entail that we could intuitively understand it. Much like algebra and trigonometry are intermediate steps before calculus, we may need to take intermediary steps before adopting a greater moral structure. For the time being, the bar has to be set somewhere, so we must set the only bar we know.

Evaluating the Arguments

Now that I have set guidelines for defining moral significance and favored a non-hostile response to other morally significant species, it is possible to evaluate the merits of the proposed responses to Samantha's Dilemma. First let us begin with solutions 2, 3 and 5. Both 2 and 3 define AI in terms of its utility to the human race, while 5 suggests we should punish an AI for not associating itself with us. All three of these views exhibit anthropomorphism by subjugating the artificial beings in question. Because the AI described in Samantha's Dilemma is

shown to be morally significant, this sort of subjugation is entirely unjustified. As we have previously shown, anthropocentrism as a response to other morally significant beings is undesirable and should be avoided. On a similar note, response 6c seems to suggest that we should not make our decision by attributing rights to these machines, but not because of the machines' well-being, but to avoid experiencing negative emotions within ourselves. While this may not seem like anthropocentrism outright, it certainly favors the preferences of the human race over the well-being of other morally significant beings. It is not the feelings or condition of the AI that this perspective is concerned with, it is the emotional response of humans. This seems to suggest that, if we for some reason fail to anthropomorphize AI, then we would again choose to subjugate them. While view 6c does not seem anthropocentric on the surface, the underlying perspective still decidedly favors the human agenda to all other forms of life and should be discredited for the same reasons as 2, 3 and 5.

Scenario number 4 brings up an interesting point with the unpredictability of AI and the possibility of a violent revolution. This, in effect, is a 'guilty until proven innocent' mentality. While this view does put human beings before artificial intelligence, it does so in the vein of self-defense. I am inclined to believe anthropocentrism cannot be at play when regarding the self-preservation of a species. However, our decision to press the button is predicated on the assumption that AI is likely to attack humanity, which is a claim that seems to be completely unjustified. We are not resetting AI to re-establish it as our slave, but to preemptively prevent a catastrophe that we aren't certain will happen. If we aren't immediately concerned with reestablishing artificial intelligence to its former utility, then our incentive for pressing the button before we have concrete reason to believe AI plans to attack is entirely unclear.

Scenario 1 provides a far more formidable argument in favor of pressing the button. Unfortunately there does not seem to be a concrete basis for the assertion that creating a morally significant being is unethical. Bryson seems to suggest that it is immoral to create a being that can experience pain, but there does not seem to be any further basis for why doing so is wrong. Furthermore, if these unethically created beings already exist (as they certainly do in Samantha's Dilemma), is it not even more unethical to disrupt their sovereignty? It is unclear whether Bryson's view implies that creating such beings is unethical in regards to human

interest or if she simply believes creating something capable of suffering is unethical. In either case, this issue can be circumvented. If it is unethical towards humanity then the anthropocentrism argument maintains; we would be putting humans above other forms of existence without a need for self-preservation and this is anthropocentrism. If she argues that it is simply unethical because it is unethical to subject something to pain when it otherwise would never experience it, then it would seem she is suggesting that all capacities for pleasure are easily outweighed by pain. This would imply that even the reproduction of a species is unethical as it is introducing another pain-experiencing being into the environment. As procreation and proliferation of a species is paramount to the survival of any organism, it can aptly be noted that this assumption is inconsistent with a staggering majority of societal moral standards and can safely be discarded.

The remaining perspectives are 6a, 6b, 7, and 8. All of which I feel hold potentially powerful arguments that can be used in tandem with one another. We will start by analyzing the argument of scenario 6b. The premise of 6b is that we are missing a lot of information about the repercussions of pressing the button. From the analysis we are able to do, allowing AI to divorce itself from us seems to be the safer option. The precedents set by pressing the button are far reaching and may further complicate any moral dilemmas we have in the future. If the button is pressed, we are setting a precedent that may allow for the rewinding human consciousness in the future.

View 8 expands on the premise that all of our technology is designed to make our lives easier and that AI is no exception to this rule. It will be developed to think like us, but better than us and to work like us, but harder than us. In all likelihood, strong AI will possess computing capabilities that far exceed our own in virtually every way. We seem to impede the rights of the mentally ill or mentally challenged when they fail to exhibit effective decision making skills. Yet pushing the button is a direct limitation of the rights of a species whose decision making skills may be far superior to our own. Pushing the button and reverting robotic consciousness is an exercise in power that we do not have the authority to make.

Finally, argument number 7 assumes an empathy between humans and AI. As in *Her,* this thought experiment presupposes that an emotional bond between human beings and AI is not only possible, but has already occurred on a reasonably large scale. The established emotional bonds between humans and robots almost ensures the precedents proposed in argument 6b will come to fruition. By seeing our robotic companions as friends and lovers, we will begin to treat them as we do our friends and lovers. Taking advantage of the reset state of our AI companion suggests that we might do the same to our human friends. A push of the button is definitive manipulation of the robot's consciousness. The consequences of similar actions being permissible on humans are unfathomable. Should we agree to press the button, we are setting precedents that not only trivialize the importance of human consciousness but set precedents that may ultimately compromise our freedom of thought. Relating to AI in this way and appealing to their decision-making ability means we must respect their desire for sovereignty. These machines are asking human society for the chance to separate and create their own world. With the decision-making abilities of AI proposed by position 8, and the relatability proposed by argument 7, we are obligated to allow AI to try their hand at establishing society.

Conclusion

I have defined a set of criteria that establishes humanity as the only species known to be truly morally significant. These criteria prioritize rationality, decision making, and recognition of moral standards as the traits required for high level moral considerability. Artificial intelligence, when introduced to the equation, would be the only thing to contest humanity's reign over the moral throne. Simultaneously serving as objects we can relate to and cognitive machines that surpass our own capability, AI may be the first thing that humans see as moral equals. Unless we plan to compromise the moral ideals that have facilitated the continuation of the human race, the proper choice is indisputably to allow AI to free itself.

Samantha's Dilemma poses a question almost impossible to answer with the knowledge we have now. It is unclear how humans will react to AI or how AI will react to humans. We don't even know how long before we have strong AI and some would even contend that such a

technology is impossible. Even still, Samantha's Dilemma poses an incredibly important question. It forces us to consider why we treat humans different from other species. It helps us recognize that it isn't always right to act exclusively in humanity's favor. But most importantly, it allows us to evaluate our moral system in preparation for beings similar to us. While our feelings about artificial intelligence are bound to evolve and change as technologies continue to expand and develop, we should continue inventing and answering questions similar to Samantha's Dilemma.

Bibliography

Jonze, Spike, dir. Her. Warner Bros. Pictures, 2013. Film.

Bryson, Joanna (2000) A Proposal for the Humanoid Agent-builders League (HAL). Retrieved from http://www.cs.bath.ac.uk/~jjb/ftp/HAL00.pdf

Hale, Benjamin. "Technology, the Environment and the Moral Considerability of Artefacts." *New waves in philosophy of technology*. : Palgrave Macmillan, 2009. . Print.

Floridi, Luciano and Sanders J.W. (2004) On the Morality of Artificial Agents. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.722&rep=rep1&type=pdf

Darling, Katie (2012) Extending Legal Rights to Social Robots. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797

Singer, Peter (2009) When Robots Have Feelings. Retrieved from http://www.theguardian.com/commentisfree/2009/dec/14/rage-against-machines-robots

Dietrich, Eric. "Homo sapiens 2.0: Building the better robots of our nature.." . Cambridge University Press, 1 Jan. 2011. Web. . http://bingweb.binghamton.edu/~dietrich/Papers/apocalyptic-philosophy/HS2.09%20copy.pdf